Learning Aligned Cross-Modal Representations from Weakly Aligned Data

Lluís Castrejón*
University of Toronto
castrejon@cs.toronto.edu

Yusuf Aytar*
MIT CSAIL
yusuf@csail.mit.edu

Carl Vondrick
MIT CSAIL
vondrick@mit.edu

Hamed Pirsiavash
University of Maryland - Baltimore County
hpirsiav@umbc.edu

Antonio Torralba
MIT CSAIL
torralba@csail.mit.edu

Abstract

People can recognize scenes across many different modalities beyond natural images. In this paper, we investigate how to learn cross-modal scene representations that transfer across modalities. To study this problem, we introduce a new cross-modal scene dataset. While convolutional neural networks can categorize cross-modal scenes well, they also learn an intermediate representation not aligned across modalities, which is undesirable for cross-modal transfer applications. We present methods to regularize cross-modal convolutional neural networks so that they have a shared representation that is agnostic of the modality. Our experiments suggest that our scene representation can help transfer representations across modalities for retrieval. Moreover, our visualizations suggest that units emerge in the shared representation that tend to activate on consistent concepts independently of the modality.

*denotes equal contribution

1. Introduction

Can you recognize the scenes in Figure 1, even though they are depicted in different modalities? Most people have the capability to perceive a concept in one modality, but represent it independently of the modality. This cross-modal ability enables people to perform some important abstraction tasks, such as learning in different modalities (cartoons, stories) and applying them in the real-world.

Unfortunately, representations in computer vision do not yet have this cross-modal capability. Standard approaches typically learn a separate representation for each modality, which works well when operating within the same modality. However, the representations learned are not aligned across modalities, which makes cross-modal transfer difficult.

Two modalities are strongly aligned if, for two images from each modality, we have correspondence at the level of objects. In contrast, weak alignment is if we only have global label that is shared across both images. For instance, if we have a picture of a bedroom and a line drawing of a

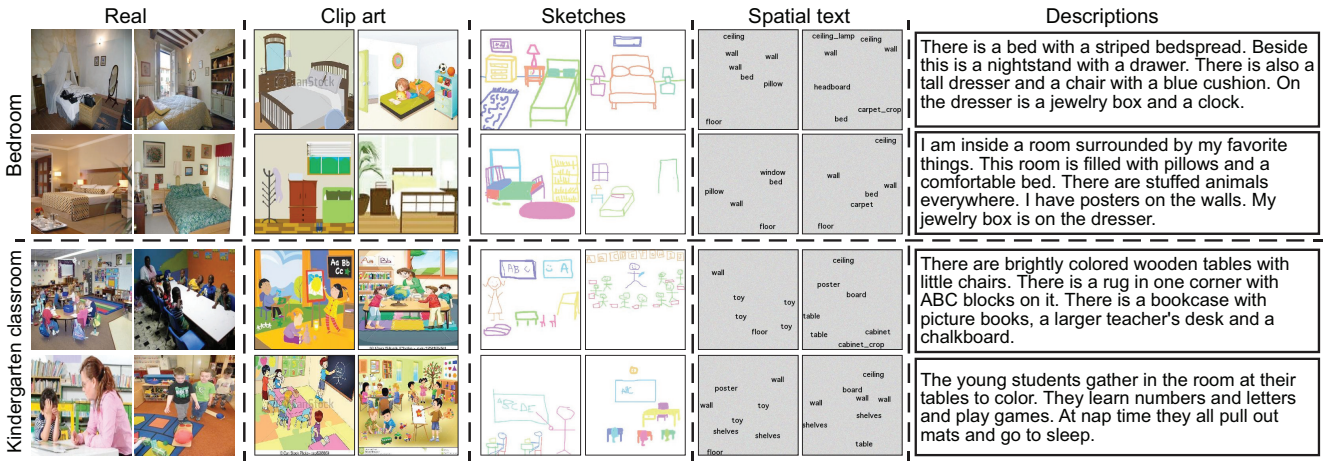


Figure 1: **Can you recognize scenes across different modalities?** Above, we show a few examples of our new cross-modal scene dataset. In this paper, we investigate how to learn cross-modal scene representations.

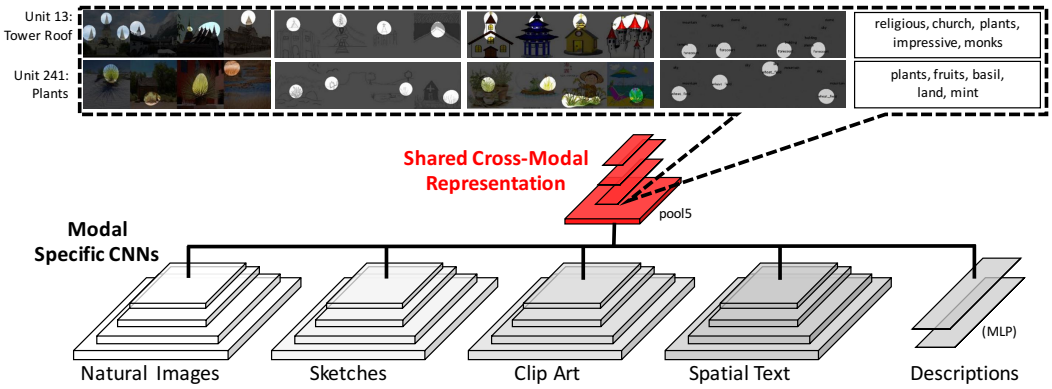


Figure 2: We learn **low-level representations** specific for each modality (white and grays) and a **high-level representation** that is shared across all modalities (red). Above, we also show masks of inputs that activate specific units the most [44]. Interestingly, although the network is trained without aligned data, units emerge in the shared representation that tend to fire on the same objects independently of the modality.

different bedroom, the only thing that we know is shared across these two images is the scene type. However, they will differ in the objects and viewpoint inside.

In this paper, our goal is to learn a representation for scenes that has strong alignment using only data with weak alignment. We seek to learn representations that will connect objects (such as bed, car) across modalities (e.g., a picture of a car, a line drawing of a car, and the word "car") without ever specifying that such a correspondence exists.

To investigate this, we assembled a new cross-modal scene dataset, which captures hundreds of natural scene types in five different modalities, and we show a few examples in Figure 1. Using this dataset and only annotations of scene categories, we propose to learn an aligned cross-modal scene representation.

We present two approaches to regularize cross-modal convolutional networks so that the intermediate representations are aligned across modalities, even when only weak alignment of scene categories is available during training. Figure 2 visualizes the representation that our full method learns. Notice that our approach learns hidden units that activate on the same object, regardless of the modality. Although the only supervision is the scene category, our approach enables alignment to emerge automatically.

Our approach builds on a foundation of domain adaptation [35, 16] and multi-modal learning [13, 29, 36] methods in computer vision. However, our focus is learning cross-modal representations when the modalities are significantly different (e.g., text and natural images) and with minimal supervision. In our approach, the only supervision we give is the scene category, and no alignments nor correspondences are annotated. To our knowledge, the adaptation of intermediate representations across several extremely different modalities with minimal supervision has not yet been extensively explored.

We believe cross-modal representations can have a large impact on several computer vision applications. For example, data in one modality may be difficult to acquire for privacy, legal, or logistic reasons (eg, images in hospitals), but may be abundant in other modalities, allowing us to train models using accessible modalities. In search, users may wish to retrieve similar natural images given a query in a modality that is simpler for a human to produce (eg, drawing or writing). Additionally, some modalities may be more effective for human-machine communication.

The remainder of this paper describes and analyzes our cross-modal representations in detail. In section 2, we first discuss related work that our work builds upon. In section 3, we introduce our new cross-modal scene dataset. In section 4, we present two complementary approaches to regularize convolutional networks so that intermediate representations are aligned across modalities. In section 5, we present several visualizations and experiments in cross-modal retrieval to evaluate our representations.

2. Related Work

Domain Adaptation: Domain adaptation techniques address the problem of learning models on some *source* data distribution that generalize to a different *target* distribution. [35] proposes a method for domain adaptation using metric learning. In [16] this approach is extended to work on unsupervised settings where one does not have access to target data labels, while [38] uses deep CNNs instead. [37] shows the biases inherent in common vision datasets and [21] proposes models that remain invariant to them. [26] learns an aligned representation for domain adaptation using CNNs and the MMD metric. Our method differs from these works in that it seeks to find a cross-modal representations between highly different modalities instead of modelling close domain shifts.

One-Shot/Zero-Shot Learning: One-shot learning techniques [10] have been developed to learn classifiers from a single or a few examples, mostly by reusing classifier parameters [11], using contextual information [27, 18] or sharing part detectors [3]. In a similar fashion, zero-shot learning [25, 31, 9, 2, 40] addresses the problem of learning new classifiers without training examples in a given domain, *e.g.* by using additional knowledge in the form of textual descriptions or attributes. The goal of our method is to learn aligned representations across domains, which could be used for zero-shot learning.

Cross-modal content retrieval and multi-modal embeddings: Large unannotated image collections are difficult to explore, and retrieving content given fine-grained queries might be a difficult task. A common solution to this issue is to use query examples from a different modality in which it is easy to express a concept (such as a clip art images, text or a sketches) and then rank the images in the collection according to their similarity to the input query. Matching can be done by establishing a similarity metric between content from different domains. [8] focuses on recovering semantically related natural images to a given sketch query and [41] uses query sketches to recover 3D shapes. [19] uses an MRF of topic models to retrieve images using text, while [33] models the correlations between visual SIFT features and text hidden topic models to retrieve media across both domains. CCA [17] and variants [34] are commonly employed methods in content retrieval. Another possibility is to learn a joint embedding for images and text in which nearest neighbors are semantically related. [13, 29] learn a semantic embedding that joins representations from a CNN trained on ImageNet and distributed word representations. [22, 43] extend them to include a decoder that maps common representations to captions. [36] maps visual features to a word semantic embedding. Our method learns a joint embedding for many different modalities, including different visual domains and text. Another group of works incorporate sound as another modality [28, 30]. Our joint representation is different from previous works in that it is initially obtained from a CNN and sentence embeddings are mapped to it. Furthermore, we do not require explicit one-to-one correspondences across modalities.

Learning from Visual Abstraction: [46] introduced clipart images for visual abstraction. The idea is to learn concepts by collecting data in the abstract world rather than the natural images so that we are not affected by mistakes in mid-level recognition *e.g.* object detectors. [12] learns dynamics and [47] learns sentence phrases in this abstract world and transfer them to natural images. Our work can complement this effort by learning models in a representation space that is invariant to modality.

3. Cross-Modal Places Dataset

We assembled a new dataset[1] to train and evaluate cross-modal scene recognition models called CMPlaces. It covers five different modalities: natural images, line drawings, cartoons, text descriptions, and spatial text images. We show a few samples from these modalities in Figure 1. Each example in the dataset is annotated with a scene label. We use the same list of 205 scene categories as Places [45], which is one of the largest scene datasets available today. Hence, the examples in our dataset span a large number of natural situations. Note that the examples in our dataset are not paired between modalities since our goal is to learn strong alignments from weakly aligned data. Furthermore, this design decision eased data collection.

We chose these modalities for two reasons. Firstly, since our goal is to study transfer across significantly different modalities, we seek modalities with different statistics to those of natural images (such as line drawings and text). Secondly, these modalities are easier to generate than real images, which is relevant to applications such as image retrieval. For each modality we select 10 random examples in each of the 205 categories for the validation set and the rest for the training set, except for natural images for which we employ the training and validation splits from [45] containing 2.5 million images.

Natural Images: We use images from the Places 205 Database [45] to form the natural images modality.

Line Drawings: We collected a new database of sketches organized into the same 205 scene categories through Amazon Mechanical Turk (AMT). The workers were presented with the WordNet description of a scene and were asked to draw it with their mouse. We instructed workers to not write text that identifies the scene (such as a sign). We collected 6,644 training examples and 2,050 validation examples.

Descriptions: We also built a database of scene descriptions through AMT. We once again presented users with the WordNet definition of a scene, but instead we asked them to write a detailed description of the scene that comes to their mind after reading the definition. We specifically asked the users to avoid using trivial words that could easily give away the scene category (such as writing "this is a bedroom"), and we encouraged them to write full paragraphs. We split our dataset into 4,307 training descriptions and 2,050 validation descriptions. We believe *Descriptions* is a good modality to study as humans communicate easily in this modality and allows to depict scenes with great detail, making it an interesting but challenging modality to transfer between.

Clip Art: We assembled a dataset of clip art images for the 205 scene categories defined in Places205. Clip art im-

[1]Dataset will be made available at http://projects.csail.mit.edu/cmplaces/

ages were collected from image search engines by using queries containing the scene category and then manually filtered. This dataset complements other cartoon datasets [46], but focuses on scenes. We believe clip art can be an interesting modality because they are readily available on the Internet and depict everyday situations. We split the dataset into 11,372 training and 1,954 validation images (some categories had less than 10 examples).

Spatial Text: Finally, we created a dataset that combines images and text. This modality consists of an image with words written on it that correspond to spatial locations of objects. We automatically construct this dataset using images from SUN [42] and its annotated objects. We created 456,300 training images and 2,050 validation images. This modality has an interesting application for content retrieval. By learning a cross-modal representation with this modality, users could use a user interface to write the names of objects and place them in the image where they want them to appear. Then, this query can be used to retrieve a natural image with a similar object layout.

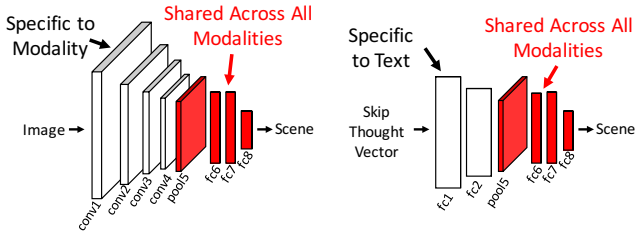
4. Cross-Modal Scene Representation

In this section we describe our approach for learning cross-modal scene representations. Our goal is to learn a strongly aligned representation for the different modalities in CMPlaces, that is, to learn a representation in which different scene parts or concepts are represented independently of the modality.for scenes that is shared across all modalities. This task is challenging partly because our training data is only annotated with scene labels instead of having one-to-one correspondences, meaning that our approach must learn a strong alignment from weakly aligned data.

4.1. Scene Networks

We extend single-modality classification networks [24] in order to handle multiple modalities. The main modifications we introduce are that we a) have one network for each modality and b) enforce higher-level layers to be shared across all modalities. The motivation is to let early layers specialize to modality specific features (such as edges in natural images, shapes in line drawings, or phrases in text), while higher layers are meant to capture higher-level concepts (such as objects) in a representation that is independent of the modality .

We show this network topology in Figure 3 with modal-specific layers (white) and shared layers (red). The modal-specific layers each produce a convolutional feature map (`pool5`), which is then fed into the shared layers (`fc6` and `fc7`). For **visual modalities**, we use the same convolutional network architecture (Figure 3a), but different weights across modalities. However, since **text** cannot be fed into a CNN (descriptions are not images), we instead encode each description into skip thought vectors [23] and



(a) Images (b) Descriptions

Figure 3: **Scene Networks:** We use two types of networks. a) For pixel based modalities, we use a CNN based off [45] to produce pool5. b) When the input is a description, we use an MLP on skip-thought vectors [23] to produce pool5 (as text cannot be easily fed into the same CNN).

use a multiple layer perceptron to map them into a representation with the same dimensionaly as `pool5` (Figure 3b). Note that, in contrast to siamese networks [5], our architecture allows learning alignments without paired data.

We could train these networks jointly end-to-end to categorize the scene label while sharing weights across modalities in higher layers. Unfortunately, we empirically discovered that this method by itself does not learn a robust cross-modal representation. This approach encourages units in the later layers to emerge that are specific to a modality (e.g., fires only on cartoon cars). Instead, our goal is to have a representation that is independent the modality (e.g., fires on cars in all modalities).

In the rest of this section, we address this problem with two complementary ideas. Our first idea modifies the popular fine-tuning procedure, but applies it on modalities instead. Our second idea is to regularize the activations in the network to have common statistics. We finally discuss how these methods can be combined.

4.2. Method A: Modality Tuning

Our first approach is inspired by finetuning, which is a popular method for transfer learning with deep architectures [6, 15, 45]. The conventional approach for finetuning is to replace the last layer of the network with a new layer for the target task. The intuition behind fine-tuning is that the earlier layers can be shared across all vision tasks (which may be difficult to learn otherwise without large amounts of data in the target task), while the later layers can specialize to the target task.

We propose a modification to the fine-tuning procedure for cross-modal alignment. Rather than replacing the last layers of the network (which are task specific), we can instead replace the earlier layers of the network (which are modality specific). By freezing the later layers in the network, we transfer a high level representation to other modalities. This approach can be viewed as finetuning the net-

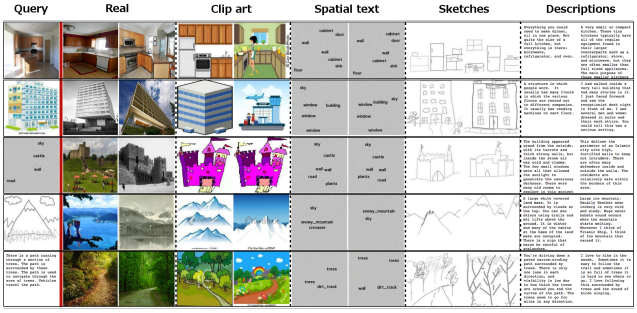


Figure 4: **Cross-Modality Retrieval :** An example of cross-modal retrieval given a query from each of the modalities. For each row, the leftmost column depicts the query example, while the rest of the columns show the top 2 ranked results in each modalitiy.

work for a modality rather than a task.

To do this, we must first learn a source representation that will be utilized for all five modalities. We use the Places-CNN network as our initial representation. Places is a reasonable representation to start with because [44] shows that high-level concepts (objects) emerge in the later layers. We then train each modal-specific network to categorize scenes in its modality *while holding the shared higher layers fixed*. Consequently, each network will be forced to produce an aligned intermediate representation so that the higher layers will categorize the correct scene.

Since the higher level layers were originally trained with only one modality (Natural images), they did not have a chance to adapt to the other modalities. After we train the networks for each modality for a fixed number of iterations, we can unfreeze the later layers, and train the full network jointly, allowing the later layers to accommodate information from the other modalities without overfitting to modal-specific representations.

Our approach is a form of curriculum learning [4]. If we train this multi-modal network with the later layers unfrozen from the beginning, units tend to specialize to a particular modality, which is undesirable for cross-modal transfer. By enforcing a curriculum to learn high level concepts first, then transfer to modalities, we can obtain representations that are more modality-invariant.

4.3. Method B: Statistical Regularization

Our second approach is to encourage intermediate layers to have similar statistics across modalities. Our approach builds upon [14, 1] who transfer statistical properties across object detection tasks. Here, we instead transfer statistical properties of the activations across modalities.

Let x_n and y_n be a training image and the scene label respectively, which we use to learn the network parameters w. We write $h_i(x_n; w)$ to refer to the hidden activations for the ith layer given input x_n, and $z(x_n; w)$ is the output of the network. During learning, we add a regularization term over hidden activations h:

$$\min_w \sum_n \mathcal{L}(z(x_n; w), y_n) + \sum_{n,i} \lambda_i \cdot \mathcal{R}_i\left(h_i(x_n; w)\right) \quad (1)$$

where the first term $\mathcal{L}$ is the standard softmax objective and the second term $\mathcal{R}$ is a regularization over the activations.[2] The importance of this regularization is controlled by the hyperparameter $\lambda_i \in \mathbb{R}$.

The purpose of $\mathcal{R}$ is to encourage activations in the intermediate hidden layers to have similar statistics across modalities. Let $P_i(h)$ be a distribution over the hidden activations in layer i. We then define $\mathcal{R}$ to be the negative log likelihood:

$$\mathcal{R}_i(h) = -\log P_i(h; \theta_i) \quad (2)$$

Since P_i is unknown we learn it by assuming it is a parametric distribution and estimating its parameters with a large training set. To that goal, we use activations in the hidden layers of Places-CNN to estimate P_i for each layer. The only constraint on P_i is that its log likelihood is differentiable with respect to h_i, as during learning we will optimize Eqn.1 via backpropagation. While there are a variety of distributions we could use, we explore two:

Multivariate Gaussian (B-Single). We consider modeling P_i with a normal distribution: $P_i(h; \mu, \Sigma) \sim \mathcal{N}(\mu, \Sigma)$. By taking the negative log likelihood, we obtain the regularization term $\mathcal{R}_i(h)$ for this choice of distribution:

$$\mathcal{R}_i(h; \mu_i, \Sigma_i) = \frac{1}{2}(h - \mu_i)^T {\Sigma_i}^{-1}(h - \mu_i) \quad (3)$$

where we have omitted a constant term that does not affect the fixed point dynamics of the objective. Notice that the derivatives $\frac{\delta \mathcal{R}_i}{\delta h}$ can be easily computed, allowing us to back-propagate this cost through the network.

Gaussian Mixture (B-GMM). We also consider using a mixture of Gaussians to parametrize P_i, which is more flexible than a single Gaussian distribution. Under this model, the negative log likelihood is:

$$\mathcal{R}_i(h; \alpha, \mu, \Sigma) = -\log \sum_{k=1}^{K} \alpha_k \cdot P_k(h; \mu_k, \Sigma_k) \quad (4)$$

such that $P_k(h; \mu, \Sigma) \sim \mathcal{N}(\mu, \Sigma)$ and $\sum_k \alpha_k = 1, \quad \alpha_k \geq 0 \ \forall_k$. Note that we have dropped the layer subscript i for clarity, however it is present on all parameters. Since $\frac{\delta \mathcal{R}_i}{\delta h}$ can be analytically computed, we can efficiently incorporate this cost into our objective during learning with backpropagation. To reduce the number of parameters, we assume the covariances Σ_k are diagonal.

[2] We omitted the weight decay from the objective for clarity. In practice, we also use weight decay.

Cross Modal Retrieval	Query	NAT				CLP				SPT				LDR				DSC				Mean
	Target	CLP	SPT	LDR	DSC	NAT	SPT	LDR	DSC	NAT	CLP	LDR	DSC	NAT	CLP	SPT	DSC	NAT	CLP	SPT	LDR	mAP
BL-Ind		17.8	15.5	10.1	0.8	11.4	13.1	9.0	0.8	9.0	10.1	5.6	0.8	4.9	7.6	6.8	0.8	0.6	0.9	0.9	0.9	6.4
BL-ShFinal		10.3	13.5	4.0	12.7	7.2	8.7	2.8	8.2	8.1	5.7	2.2	9.3	2.4	2.5	3.1	3.2	3.3	3.4	8.5	2.4	6.1
BL-ShAll		15.9	14.2	9.1	0.8	8.9	10.9	7.0	0.8	8.4	7.4	4.2	0.8	4.3	5.6	5.7	0.8	0.6	0.9	0.9	0.9	5.4
A: Tune		12.9	23.5	5.8	19.6	9.7	15.5	4.0	13.7	19.0	13.5	5.6	24.0	4.1	3.8	5.8	5.9	6.4	4.5	9.5	2.5	10.5
A: Tune (Free)		14.0	29.8	6.2	18.4	9.2	17.6	3.7	12.9	21.8	15.9	6.2	27.7	3.7	3.1	6.6	5.4	5.2	3.5	10.5	2.1	11.2
B: StatReg (Gaussian)		18.6	20.2	10.2	0.8	11.1	15.4	8.5	0.8	13.3	15.1	7.7	0.8	4.7	6.6	6.9	0.9	0.6	0.9	0.8	0.9	7.2
B: StatReg (GMM)		17.8	23.7	9.5	5.6	13.4	18.1	8.9	4.6	16.7	16.2	8.8	5.3	6.2	8.1	9.4	3.3	3.0	4.1	4.6	2.8	9.5
C: Tune + StatReg (GMM)		14.3	32.1	5.4	22.1	10.0	19.1	3.8	14.4	24.4	17.5	5.8	32.7	3.3	3.4	6.0	4.9	15.1	12.5	32.6	4.6	**14.2**

Table 1: **Cross-Modal Retrieval mAP:** We report the mean average precision (mAP) on retrieving images across modalities using `fc7` features. Each column shows a different query-target pair. On the far right, we average over all pairs. For comparison, chance obtains 0.73mAP. Our methods perform better on average than the finetuning baselines with method C performing the best.

Cross-Modal Retrieval vs Layers	pool5	fc6	fc7
BL-Ind	1.9	3.6	6.4
BL-ShFinal	1.5	3.0	6.1
BL-ShAll	1.8	3.1	5.4
A: Tune	4.8	10.7	10.5
A: Tune (Free)	4.4	9.9	11.2
B: StatReg (Gaussian)	2.0	4.7	7.2
B: StatReg (GMM)	2.0	7.5	9.5
C: Tune + StatReg (GMM)	3.6	13.2	14.2

Table 2: **Mean Cross-Modal Retrieval mAPs across Layers:** Note that the baseline results decrease drastically as we go lower levels (e.g. `fc7` to `fc6`) in the deep network. However the alignment approaches are much less affected.

Within Modality Retrieval	NAT	CLP	SPT	LDR	DSC	Mean
BL-Ind	19.3	31.7	83.0	18.1	11.1	32.6
BL-ShFinal	18.2	22.0	81.2	13.8	29.8	33.0
BL-ShAll	18.4	26.7	82.7	16.6	11.1	31.1
A: Tune	19.0	23.9	74.2	13.7	36.3	33.4
A: Tune (Free)	19.4	22.9	85.0	13.5	34.2	**35.0**
B: StatReg (Gaussian)	19.4	31.1	84.0	17.3	11.1	32.6
B: StatReg (GMM)	19.3	31.1	82.5	16.7	13.5	32.6
C: Tune + B: StatReg (GMM)	20.2	22.5	82.2	13.1	37.0	**35.0**

Table 3: **Within Modal Retrieval mAPs:** We report the mean average precision (mAP) for retrieving images within the same modality using `fc7` features.

We fit a separate distribution for each of the regularized layers in our experiments (`pool5`, `fc6`, `fc7`). During learning, the optimization will favor solutions that categorize the scene but also have an internal shared representation that is likely under P_i. Since P_i is estimated using Places-CNN, we are enforcing each modality network to have similar higher layers statistics to those of Places-CNN.

4.4. Method C: Joint Method

The two proposed methods (A and B) operate on complementary principles and may be jointly applied while learning the networks. We combine both methods by first fixing the shared layers for a given number of iterations. Then, we unfreeze the weights of the shared layers, but now train with the regularization of method B to encourage activations to be statistically similar across modalities and avoid overfitting to a specific modality.

4.5. Implementation Details

We implemented our network models using Caffe [20]. Both our methods build on top of the model described in [24], with the modification that the activations from layers `pool5` onwards are shared across modalities, and layers before are modal-specific. Architectures for method A only use standard layer types found in the default version of the framework. In contrast, for model B we implemented a layer to perform regularization given the statistics of a GMM as explained in the previous sections. In our experiments the GMM models are composed by $K = 100$ different single gaussians.

For each model we have a separate CNN initialized using the weights of Places-CNN [45]. The weights in the lower-layers can adapt independently for each modality, while we impose restrictions in the higher layer weights as explained for each method. Because CNNs start training from a good initialization, we set up the learning rate to $lr = 1e^{-3}$ (higher learning rates made our models diverge). We train the models using Stochastic Gradient Descent.

To adapt textual data to our models we use the network architecture described here. First, we represent descriptions by average-pooling the Skip-thought [23] representations of each sentence in a given description (a description contains multiple sentences). To adapt this input to our shared representation we employ a 2-layer MLP. The layer size is constant and equal to 4800 units, which is the same dimensionality as that of a Skip-thought vector, and we use ReLU nonlinearities. The weights of these layers are initialized using a gaussian distribution with $std = 0.1$. This choice is important as the statistics of the Skip-thought representations are quite different to those of images and inadequate weight initializations prevent the network from adapting textual descriptions to the shared representation. Finally, the output layer of the MLP is fully-connected to the first layer (`pool5`) of our shared representation.

5. Experimental Results

Our goal in this paper is to learn a representation that is aligned across modalities. We show three main results that evaluate how well our methods address this problem. First, we perform cross-modal retrieval of semantically-related content. Secondly, we show visualizations of the learned representations that give a qualitative measure of how this alignment is achieved. Finally, we show we can reconstruct natural images from other modalities using the features in the aligned representation as a qualitative measure of which semantics are preserved in our cross-modal representation.

5.1. Cross-Modality Retrieval

In this experiment we test the performance of our models to retrieve content depicting the same scene across modalities. Our hypothesis is that, if our representation is strongly aligned, then nearest neighbors in this common representation will be semantically related and similar scenes will be retrieved.

We proceed by first extracting features for the validation set of each modality from the shared layers of our cross-modal representation. Then, for every modality, we randomly sample a query image and compute the cosine distance to the extracted feature vectors of all content in the other modalities. We rank the documents according to the distances and compute the Average Precision (AP) when using the scene labels. We repeat this procedure 1000 times and report the obtained mean APs for cross-modality retrieval in Table 1 and the results for within-modality retrieval in Table 3. For completeness, we also show examples of retrievals in Figure 4. We compare our results against three different finetuning baselines:

Finetuning individual networks (BL-Ind): In this baseline we finetune a separate CNN for each of the modalities. The CNNs follow the AlexNet [24] architecture and are initialized with the weights of Places-CNN. We then finetune each one of them using the training set from the corresponding modality. This is the current standard approach employed in the computer vision community, but it does not enforce the representations in higher CNN layers to be aligned across modalities.

Finetuning with shared final layers (BL-ShFinal): similarly to our method A, we force networks for each modality to share layers from `pool5` onwards. However, as opposed to our method, in this baseline we do not fix the weights in the shared layers and instead let them be updated by backpropagation of the scene classification error.

Finetuning with a single shared CNN (BL-ShAll): here we use a single instance of Places-CNN shared by all modalities. We finetune it using batches that contain data from each modality. Note that this baseline can only be applied to pixel data because of the different architecture required for text, hence we excluded the descriptions here.

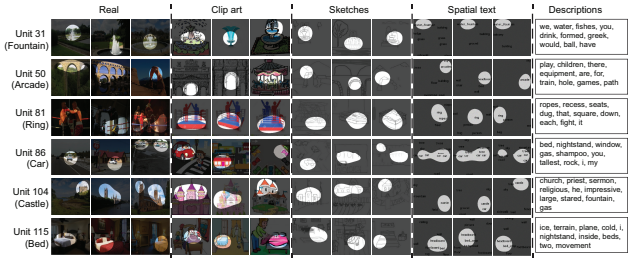


Figure 5: **Visualizing Unit Activations:** We visualize `pool5` in our cross-modal representation above by finding masks of images/descriptions that activate a specific unit the most [44]. Interestingly, the same unit learns to detect the same concept across modalities, suggesting that it may has learned to generalize across these modalities.

However, we employed the textual features from BL-Ind for completeness.

CCA approaches are common for cross-modal retrieval, however past approaches were not directly comparable to our method. Standard CCA requires sample-level alignment, which is missing in our dataset. Cluster CCA [34] works for class-level alignments, but the formulation is intended for only two modalities. On the other hand, Generalized CCA [17] does work for multiple modalities but still requires sample-level alignments. Concurrent work with ours extends CCA to multi-label settings [32].

As displayed in Table 1 both method A and B improve over all baselines, suggesting that the proposed methods have a better semantic alignment in `fc7`. Furthermore, method C outperforms all other reported methods. Particularly, we can observe how method C is able to obtain a comparable performance for retrievals using descriptions to method A, while retaining the superior performance of method B for the other modalities. Note that in our experiments the baseline methods perform similarly to our method in all modalities except for descriptions, as they were not able to align the textual and visual data very well. Also note that the performance gap between our method and the baselines increases as modalities differ from each other (see SPT and DSC results). For statistical regularization, using GMM instead of a single Gaussian also notably improves the performance, arguably because of the increased complexity of the model.

Table 3 reports the within-modal retrieval results. By performing alignment through proposed methods, we also increase within-modal retrieval results on average. Table 2 shows the mean performances across layers. We can observe how in general the proposed methods outperform the different baselines in cross-modal retrieval for each of the layers. We can also observe how, as we use features from higher layers in the CNN, the results improve, since they represent higher-level semantics closer to the scene label.

5.2. Hidden Unit Visualizations

We now investigate what input data activates units in our shared representation. For **visual data**, we use a visualization similar to [44]. For **textual descriptions**, we compute the paragraphs that maximally activate each filter, and then we employ tf-idf features to determine the most common relevant words in these paragraphs.

Figure 5 shows, for some of the 256 filters in `pool5`, the images in each visual modality that maximally activated the filter with their mask superimposed, as well as the most common words in the paragraphs that maximally activated the units. We can observe how the same concept can be detected across modalities without having explicitly aligned training data. These results suggest that our method is learning some strong alignments across modality only using weak labels coming from the scene categories.

To quantify this observation, we set up an experiment. We showed human subjects activations of 100 random units from `pool5`. These activations included the top five responses in each modality with their mask. The task was to select, for each unit, those images that depicted a common concept if it existed. Activations could be generated from either the baseline BL-Ind or from our method A, but this information is hidden from the subjects.

After running the experiment, we selected those results in which at least 4 images for the real modality were selected. This ensured that the results were not noisy and were produced using units with consistent activations, as we empirically found this to be a good indicator of whether a unit represented an aligned concept. We then computed the number of times subjects selected at least one image in each of the other modalities. With our method, 33% of the times this process selected at least one image from each modality, whereas for the baseline this only happened 25% of the times. Furthermore, 19% of the times we selected at least two images for each modality as opposed to only 14% for the baseline. These results suggest that, when a unit is detecting a clear concept, our method outperforms the best finetuning method and can strongly align the different modalities.

5.3. Feature Reconstructions

Here we investigate if we can generate images in different modalities given a query. The motivation is to gain some visual understanding of which concepts are preserved across modalities and which information is discarded [39]. We use the reconstruction approach from [7] out-of-the-box, but we train the network using our features. We learn an inverting network for each modality that learns a mapping from features in the shared `pool5` layer to downsampled reconstructions of the original images. We refer readers to [7] for full details. We employ `pool5` features as opposed to `fc7` features because the amount of compression

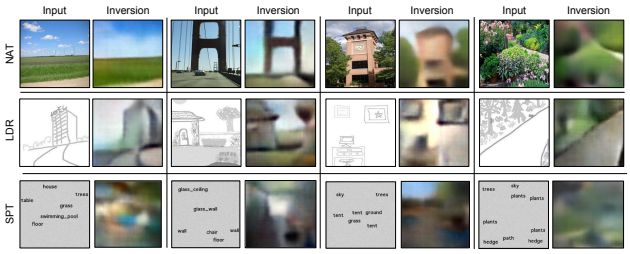


Figure 6: **Inverting features across modalities:** We visualize some of the generated images by our inverting network trained on real images. **Top row:** reconstructions from real images. These preserve most of the details of the original image but are blurry because of the low dimensionality of the `pool5` representation. **Middle row:** reconstructions from line drawings, wehere the network adds colors to the reconstructions while preserving the original scene composition. **Bottom row:** inversions from the spatial text modality. Reconstructions are less detailed but roughly preserve the location, shape and colors of the different parts of the input scene.

of the input image in the latter produces worse reconstructions.

If concepts in our representation are correctly aligned, our hypothesis is that the reconstruction network will learn to generate images that capture the statistics of the data in the output modality and while show same concepts across modalities in similar spatial locations. Note that one limitation of these inversions is that output images are blurry, even when reconstructing images within a same modality, due to the data compression in `pool5`. However, our reconstructions have similar quality to those in [7] when reconstructing from `pool5` features within a modality.

Figure 6 shows some successful examples of reconstructions. We observed this is a hard, arguably because the statistics of the activations in the common representation are very different across modalities despite the alignment, which might be due to the reduced amount of information in some of the modalities (i.e. clipart and spatial text images contain much less information that natural images). However, we note that in the examples the trained model is capable of reproducing the statistics of the output modality. Moreover, the reconstructions usually depict the same concepts present in the original image, indicating that our representation is aligning and preserving scene information across modalities.

Acknowledgements: We thank TIG for managing our computer cluster. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used for this research. This work was supported by NSF grant IIS-1524817, by a Google faculty research award to A.T and by a Google Ph.D. fellowship to C.V.

References

[1] Y. Aytar and A. Zisserman. Part level transfer regularization for enhancing exemplar svms. In *Computer Vision and Image Understanding*, 2015. 5

[2] J. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. *arXiv*, 2015. 3

[3] E. Bart and S. Ullman. Cross-generalization: Learning novel classes from a single example by feature replacement. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 672–679. IEEE, 2005. 3

[4] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009. 5

[5] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah. Signature verification using a siamese time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688, 1993. 4

[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013. 4

[7] A. Dosovitskiy and T. Brox. Inverting convolutional networks with convolutional networks. *arXiv*, 2015. 8

[8] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *TVCG*, 2011. 3

[9] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*, 2013. 3

[10] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *TPAMI*, 2006. 3

[11] M. Fink. Object classification from a single example utilizing class relevance metrics. *NIPS*, 2005. 3

[12] D. F. Fouhey and C. L. Zitnick. Predicting object dynamics in scenes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2027–2034. IEEE, 2014. 3

[13] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 2, 3

[14] T. Gao, M. Stark, and D. Koller. What makes a good detector?–structured priors for learning from few examples. In *Computer Vision–ECCV 2012*, pages 354–367. Springer, 2012. 5

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 4

[16] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011. 2

[17] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004. 3, 7

[18] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 654–661. IEEE, 2005. 3

[19] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *ICCV*, 2011. 3

[20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014. 6

[21] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba. Undoing the damage of dataset bias. In *ECCV*, 2012. 2

[22] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv*, 2014. 3

[23] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler. Skip-thought vectors. *arXiv*, 2015. 4, 6

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 4, 6, 7

[25] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 3

[26] M. Long and J. Wang. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015. 2

[27] K. Murphy and W. Freeman. Contextual models for object detection using boosted random fields. NIPS, 2004. 3

[28] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011. 3

[29] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv*, 2013. 2, 3

[30] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. *arXiv preprint arXiv:1512.08512*, 2015. 3

[31] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009. 3

[32] V. Ranjan, N. Rasiwasia, and C. Jawahar. Multi-label cross-modal retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4094–4102, 2015. 7

[33] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ICM*, 2010. 3

[34] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal. Cluster canonical correlation analysis. In *AISTATS*, pages 823–831, 2014. 3, 7

[35] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 2

[36] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013. 2, 3

[37] A. Torralba, A. Efros, et al. Unbiased look at dataset bias. In *CVPR*, 2011. 2

[38] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015. 2

[39] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles: Visualizing object detection features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–8, 2013. 8

[40] C. Vondrick, H. Pirsiavash, A. Oliva, and A. Torralba. Learning visual biases from human imagination. In *Advances in Neural Information Processing Systems*, 2015. 3

[41] F. Wang, L. Kang, and Y. Li. Sketch-based 3d shape retrieval using convolutional neural networks. *arXiv*, 2015. 3

[42] J. Xiao, J. Hays, K. Ehinger, A. Oliva, A. Torralba, et al. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 4

[43] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015. 3

[44] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. *arXiv*, 2014. 2, 5, 7, 8

[45] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014. 3, 4, 6

[46] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *CVPR*, 2013. 3, 4

[47] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1681–1688. IEEE, 2013. 3